DOCUMENT RESUME

ED 340 755 TM 018 011

AUTHOR Dunbar, Stephen B.
TITLE On the Development of a National Assessment of College Student Learning: Measurement Policy and Practice in Perspective. Draft.
SPONS AGENCY National Center for Education Statistics (ED), Washington, DC.
PUB DATE 1 Oct 91
NOTE 45p.; Commissioned paper prepared for a workshop on Assessing Higher Order Thinking & Communication Skills in College Graduates (Washington, DC, November 17-19, 1991), in support of National Education Goal V, Objective 5. For other workshop papers, see TM 018 009-024.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Academic Achievement; *College Graduates; Communication Skills; Critical Thinking; *Educational Assessment; Educational Policy; *Evaluation Methods; Evaluation Utilization; Federal Programs; Higher Education; National Programs; Problem Solving; *Program Development; Student Evaluation; Testing Programs; Thinking Skills
IDENTIFIERS America 2000; *National Assessment of College Student Learning; *National Education Goals 1990

ABSTRACT

In light of the National Education Goals of 1990, several arguments against any federally funded large-scale census approach to the assessment of college student learning in the United States are presented to clarify challenges to an objective of Goal 5, which specifies that "the proportion of college graduates who demonstrate advanced ability to think critically, communicate effectively, and solve problems will increase substantially." There are really two central measurement problems posed by the initiative. One problem is that of measuring skills in the domains of critical thinking, communication, and problem solving; and the other problem concerns the measurement of social values for educational achievement at the college level. More specific measurement issues for the National Assessment of College Student Learning (NACSL) are: (1) consequences of measurement; (2) content of measurement; and (3) setting standards and determining their stability. Because America 2000 involves many controversial areas in measurement, the development and implementation of the NACSL should proceed slowly and carefully. A research plan is proposed to create the richest possible source of data about college student learning. A 27-item list of references is included. Reviews by J. Chaffee, S. B. Dunbar, and R. K. Hambleton of this position paper are provided. (SLD)

On the Development of a National Assessment of College Student Learning: Measurement Policy and Practice in Perspective

Stephen B. Dunbar

The University of Iowa

October 1, 1991

Paper commissioned by the National Center for Educational Statistics and prepared for the Higher Order Thinking and Communication Skills Study Design Workshop, Washington, D.C., November 17-19, 1991.

It should come as no surprise to careful observers of the current assessment climate in the United States that eventually the federal government would turn its attention to the matter of assessment and accountability in postsecondary education. Despite the many obstacles that have always faced attempts to measure the impact of a college education, or even to articulate the value of a college education to a general audience, we are now faced with an initiative that poses some extremely difficult questions. Some of these questions go well beyond any experiences educators in the United States have had in the arena of assessment and public policy. Many of them present problems of a technical nature that have never been posed to specialists in measurement theory or practice. All of them require answers if the national assessment of college student learning (hereafter NACSL) is to become a component of federal educational policy.

This paper begins by taking a careful look at the impetus for the federal government's current interest in assessment at the college level and considers implications for measurement. Several arguments against any federally funded, large-scale, census approach to the assessment of college student learning in the United States are presented in order to clarify challenges for the developmental effort. Although these arguments may not always appear to reflect principles of educational measurement *per se*, they have their origins in such principles and in the expected quality and utility of the information that might derive from a census approach for measuring college student learning. They are arguments that the Education Department (ED) must weigh and

consider before proceeding with any major effort of this sort. The second section of the paper outlines specific measurement problems in higher education outcomes assessment that will require solutions -- value judgments is perhaps a more accurate phrasing because they are only solutions after one ascribes a certain set of political and social values to the assessment process -- before any effort at national assessment of college student learning can proceed. This third section advances an approach to the endeavor that is consistent with current knowledge about (1) the effects of tests as instruments of educational reform, (2) the degree to which higher-order skills transfer across subject matters and (3) the technology of setting performance standards. The final section of the paper attempts to place measurement issues in to the broader context of educational reform in American colleges and universities.

I: Goals, Scales and Measurements

The foundation of the present initiative for NACSL is an objective for Goal 5 of the educational reform strategy outlined in the *America 2000* report that states

> The proportion of college graduates who demonstrate advanced ability to think critically, communicate effectively, and solve problems will *increase substantially* (*America 2000: An Education Strategy*, p. 64, emphasis added).

Who in his/her right mind would argue with a statement such as this, one which, however generally, speaks to the heart of what Americans have come to value from higher education. Without question, our nation needs to safeguard its role

of international leadership in higher education. However, two aspects of this goal statement deserve special comment because of their consequences for measurement.

First, it is important to note that the goal statement is phrased in terms of skills that are generally understood among educated persons as generalized outcomes of higher education, but which do not have an unambiguous status in the curricula of many colleges and universities. Few colleges explicitly teach critical thinking, communicative competence, or problem-solving. True, many colleges and universities have instituted requirements in the area of general education that target these skills. However, the intent of this goal statement is not critical thinking, communication, or problem solving with respect to a specific curriculum in the area of general education. Of interest is whether or not college students can demonstrate these skills in context, on the job, for example, or in the voting booth. Generalized skills of this nature are not meaningfully taught, nor measured, in a vacuum.[1] Perhaps by definition, some degree of complexity in *content* is needed in order to exercise a desired level of complexity in the cognitive *processes* of critical thinking, communication, and problem-solving (Snow & Lohman, 1989). What to measure is not transparent from the language of the Goal 5 objective.

[1]Baird (1988) provides a useful perspective on this aspect of the target skills for NACSL. There is also a growing literature in cognitive psychology that indicates we know much more about how such skills operate within specific disciplines than we do about how they transfer across disciplines or how they function as unitary psychological constructs in their own right. See Resnick (1989) for further discussion of these issues.

Another noteworthy aspect of the *America 2000* goal statement is its focus on gain in critical thinking, communication, and problem-solving skills, gain of some *unspecified* magnitude that in the judgment of, say, some *unspecified* blue-ribbon panel, is an amount that will prepare an educated citizenry to be competitive in a global economy of dwindling resources. It is this kind of scale for achievement in higher education, a scale ultimately defined on the basis of the *utility* of higher education in realizing certain desirable social goals, rather than one defined on the basis of observations of specific learning objectives, that is understood to be the driving force behind the present initiative for NACSL. **Thus, there are actually two scales of interest that need to be defined for NACSL. The first is an achievement scale and the second is a social utility scale, and the inferences that truly respond to the goal statement of *America 2000* require that the achievement scale be mapped onto the utility scale.**

The fact that the goal of interest concerns skills that are not directly taught, coupled with the fact that the ultimate scale of interest is based on social utility and not on achievement *per se*, presents some measurement problems that do not exist in other assessments of educational progress that have been used at local, state, and national levels. Achievements that are, by their very nature, meaningful to the public in the context of a real-life setting do not lend themselves well to observation and objective measurement. Moreover, surrogates for them are prone to charges of contrivance, and the data from them don't generalize well to the criterion situations of real interest (Dunbar, Koretz & Hoover, 1991).

Further, it would be naive to assume that scores from either an off-the-shelf test or one developed specifically for the purpose of NACSL could, by themselves, provide the information that is needed to render a judgment of value about attainment of the goal. Presumably, a governmentally appointed body will have to determine whether observed gains, however they happen to be measured, are "substantial." The recent controversy surrounding the attempt by the National Assessment Governing Board (NAGB) to set this kind of standard for a comparatively well-defined domain in the NAEP Mathematics Assessment (Rothman, 1991) might well pale in comparison to the disagreements likely to arise in determining whether or not gains in, say, critical thinking skills are substantial enough. There is no value-free and thereby objective scale for the kind of measurement that the *America 2000* envisions for higher education in the United States. Moreover, any scale that is developed for NACSL will ultimately be interpreted with respect to some such scale of social utility, because that is the real foundation of the goal statement from *America 2000*, whether we like it or not.

From a broad perspective, then, there are really two central measurement problems posed by the NACSL initiative. The more tractable of the two is the problem of measuring a particular set of skills in the domains of critical thinking, communication, and problem-solving, tractable, as will be argued, to the extent that a genuine consensus can be reached about what these skills consist in. Traditional instrument development procedures that entail the delineation of content domains, their translation into tables of specifications, and the writing of

performance tasks and test questions can provide a place to start solving the first problem.

The other measurement problem, buried in the language of the *America 2000* report, concerns the measurement of social values for educational achievement at the college level.[2] This aspect of the NACSL initiative sets it apart from other federal assessment efforts such as NAEP. NAEP began with a very modest purpose, to provide descriptive information to the public concerning educational progress, and it responded to the charge with simple reports that emphasized concrete examples of exercises and results, much like a public opinion poll would do. Like the Gross National product or the Consumer Price Index, NAEP provided the data, but wasn't itself a policy instrument. Only over time has NAEP, or perhaps more precisely NAGB, experimented with the added political responsibility of judging the adequacy of achievement by setting standards that in effect attempt to map the achievement scale onto the utility scale. NACSL, on the other hand, if it is to be proposed as

---

[2] At one time, there was some agreement on methods that might be reasonably used to develop psychological scales for judgments of value (cf. Thurstone, 1959). But the social landscape in the United States has changed dramatically since that time. It is no longer the case that a self-respecting measurement psychologist would presume to possess a storehouse of procedures that could be used to map an achievement scale onto a universal scale of social utility that could withstand rightful challenges to it in a diverse society such as ours. Standard setting procedures (Angoff, 1971; Nedelsky, 1954; Ebel, 1956), which have been used in education to create the kind of scale at issue, have always had their critics (see, for example, Glass, 1976, and others in the same volume; Jaeger, 1989). These critics can be expected to become quite vocal when the requisite inference for a national effort hinges on whether performance is "good enough."

an endeavor that responds directly to the *America 2000* goal for college students, should be recognized from the outset as having been charged with a judgmental responsibility.[3] That responsibility ought to shape development efforts for NACSL to a considerable degree.

II. Measurement Issues for NACSL

No attempt will be made in this section to review the literature on critical thinking, communication, and problem-solving and their measurement. In spite of the fact that they are universally associated with higher learning, we don't usually think of them as discipline-based; rather they are part of the tacit curriculum of any major field in college. English majors become accomplished when they master the rhetoric of literary criticism, mathematics majors when they can prove theorems by induction, chemistry majors when they design experiments to isolate the elements of an unknown compound, and music majors when they present an original interpretation of a piece of classical music. The extent to which an accomplished math major will exercise critical thinking or problem-solving skills in real-life settings that don't involve expertise in mathematics is uncertain. **All of the arguments in this section are based on the premise that the criterion situations of interest in response to Goal 5 of *America 2000* involve the target skills *in context*, and that any context-free**

[3] By this statement, no suggestion is being made that NACSL should be conceived in this way. On the contrary, the main purpose behind these arguments is to show that the goal statement is unreasonable as a charge for the development of NACSL, precisely because it establishes a utility scale which is beyond the reach of any technical procedures for standard setting. If NACSL is itself expected to support inferences of the kind reflected by the goal statement, it will not be beyond reproach.

surrogate developed in the interest of efficiency and expediency will not only fail to satisfy the audience for NACSL but will also suffer from major shortcomings with respect to validity, broadly defined.

Validity Issues for NACSL: Consequences of Measurement

The most recent considerations of validity in educational measurement define the concept broadly to include both evidential and consequential bases for the interpretations that are made of test scores (Messick, 1989). In the present way of thinking, the use of an assessment instrument can no longer be considered valid simply because the content of the test items or performance tasks matches a table of specifications developed by a consensus panel, or because the scores have reasonably high correlations with other measures of the target skills or with job performance. This kind of evidence (traditionally termed content, criterion-related, and construct validity evidence) is considered necessary for a validity argument, but it is no longer accepted by specialists as sufficient for validity. The *consequences* of assessment must form an integral part of the justification for anticipated uses of a particular instrument, or perhaps in the case of NACSL, a system of instruments.

This new emphasis on the consequences of measurement focuses attention on the importance of well articulated purposes behind the endeavor. The only purpose that can be identified for NACSL at the present time has to do with meeting the challenge posed by *America 2000*. The apparent purpose of *America 2000* is to provide the impetus for broadly-based reform efforts at all levels of education in the United States, again the social utility notion.

Unfortunately, the use of tests of whatever kind to create an atmosphere hospitable to educational reform is an area of great uncertainty. It is a use of assessment about which measurement specialists know the least vis-a-vis consequential validity. We have no experiments to fall back on, only a history of experiences, largely negative experiences, from the use of tests to drive educational reform in the public schools.

There are many examples of controversial uses of tests in shaping public policy in the schools. The first glimpse of experiences yet to come was provided by the minimum-competency testing (MCT) movement in the 1970s. MCT programs in states like Kansas and Florida were designed as gatekeepers for either grade-to-grade promotion or high school graduation, and were intended to ensure that no one failing to satisfy basic standards of achievement would be promoted or graduated. Minimum standards were, and continue to be (cf. Gallup & Elam, 1988) in the public's eye implicitly amenable to measurement by objective tests. The intent of such standards from the very beginning was to improve education; however, their implementation through mandatory requirements for testing had several undesirable results.

The public's implicit trust in tests to measure basic competencies only temporarily masked some of the consequences of basing a decision of enormous social import on a single score. The negative consequences were probably not fully appreciated, if they are so today, until after a class action suit was filed in U. S. District Court in Florida, which we know as Debra P. v Turlington. A discrepancy of nearly 20 percentage points between the failure

rates of blacks and whites on a functional literacy test used to deny high school diplomas focused national attention on the impact of high-stakes testing on individuals and schools. Consequential validity became the crux of the legal arguments, pro and con, concerning test use in Debra P. v Turlington. In addition, terms that had gone out of vogue in the measurement literature decades before, curricular and instructional validity for example, were resurrected in an effort to cope with judicial rulings in the Debra P. case. These concepts remain with us today as important dimensions of consequential validity. They show the importance of establishing rigorous standards for validity in NACSL long before the first press release telling the nation how its college students are doing.

In the aftermath of the Debra P. case, states approached the task of using tests to make decisions about individual students with much greater attention to questions of consequential validity. However, the Carnegie Commission reports of the early 1980s gave states and local school districts a different kind of motive for increasing their testing activities. It was a motive that extended into the halls of academe as well. In the test-based reform movement of the 1980s, tests were not, by and large, advocated because of what they could tell us about individuals, but instead because of the information they could provide about an aggregate group, a school building, for example, an academic department, a local district, or an entire state. Using tests to estimate the educational achievement of groups, rather than individuals, at the time was probably considered safe and justifiable practice in terms of consequences because the welfare of individual students was not at risk. In retrospect, however, such uses

of tests were indeed high-stakes because, at least in the public schools, they often had the effect of focusing the attention of teachers on the more narrowly defined content of a test to which they would be held accountable rather than on the broader domains of achievement from which the test specifications were originally derived. Teaching to the test, not to the domain, has been a common criticism of test-driven reform efforts (Shepard, 1990).

Although the effects of the test-based reform movement of the 1980s is currently an area of intense debate in the measurement and public policy communities, evidence is beginning to be assembled that leads to some general conclusions about consequences. The well-known Lake Wobegon effect (Cannell, 1989; Linn, Graue & Sanders, 1990) is generally accepted as one consequence of test-based accountability programs. In such programs, instruction can become so narrowly focused on the content of particular tests that the national norms for such tests are invalidated and make it appear that all students are above average. Preliminary findings from another study of the effects of high-stakes testing programs (Koretz, Linn, Dunbar & Shepard, 1991) indicate that performance on tests used in such situations does not generalize well even to performance on highly similar measures of the same content domains. In high-stakes testing situations studied by Koretz, et al, standardized achievement tests other than the one used for school district accountability yielded generally lower scores than the ones that were reported for accountability purposes. The implication of this finding is that when the stakes for testing are high, test scores may give an overly optimistic picture of actual achievement.

These experiences illustrate some potential concerns about the consequences of an assessment that has as a central purpose the need to document educational reform. The ultimate consequences of an assessment predicated on reform are not easily predicted. Nevertheless, the range of potential consequences must be duly considered in the design of any assessment, especially one that will become the focus of considerable national attention. The approach to the development of NACSL described in the next section of this paper is intended to carefully guard against unintended consequences.

Validity Issues for NACSL: Content of Measurement

As mentioned previously, the content domains identified in Goal 5 for the assessment of college student learning are difficult to operationalize for purposes of measurement. The fact that they are critical domains for general education programs at American colleges and universities doesn't make them more tractable. General education, as a curriculum in higher education and as a construct for educational measurement, simply lacks the focus that is truly necessary for the development of instruments that are content valid in the judgment of experts (Yarbrough, 1991). Existing measures of the outcomes of general education (e.g. ACT-COMP, CAAP, and Academic Profiles) have been the subject of much criticism in the literature on assessment in higher education as much because of the ill-defined nature of the domain as because of internal limitations of the measures themselves.

Even if it were the case that a consensus could be reached on the constituent skills that relate to critical thinking and problem solving in general, for example if a universal general education curriculum existed at all American colleges and universities, there is still a body of research in cognitive psychology that portends difficulties with respect to validity. Given that the inferences to be made from NACSL entail successful transfer between the tasks used for assessment and the real-world criterion situations that motivate the *America 2000* ideal, a natural question from a content/construct view of validity asks about the degree of transfer that can be expected from general assessments of higher-order skills to particular applications at some later point in time. Larkin (1989) traced developments in cognitive psychology on the question of transfer and concluded

> Although attractive, the notion that transferable knowledge is a core of general problem-solving skills has been historically unproductive. There is not good evidence that instruction in such skills improves performance . . . There is evidence of varying kinds and of varying strengths that skills that are somewhat domain specific may transfer. These include strategies that apply to a moderately broad range of domains, skills for managing the surroundings of a task, and skills for learning. None of these kinds of knowledge, however, forms a complete routine that can be executed in the absence of other knowledge; all are intermingled with other domain-specific knowledge (Larkin, 1989, pp. 303-4).

While it is probably misleading to say that the above point of view is universally accepted as a matter of fact among cognitive psychologists, the recognition of uncertainty with regard to transfer of higher-order skills is characteristic of writers in the field. Snow and Lohman (1989), for example, suggest that "much

important cognitive activity is domain specific" and that "it seems likely that even general problem-solving strategies are conditioned by or adapted to the particular characteristics of the knowledge domain in which they are used" (p. 305). If these writers are correct, then from a measurement point of view, it is unlikely that any instrument that proposes to measure critical thinking, communication, and problem-solving as general intellectual skills[4] is not likely to satisfy accepted standards for the content and construct validity of educational measurements, particularly those that are used for external assessments of educational progress.

Setting Standards for NACSL: Measurement and Values

The most challenging task facing NCES given the language of *America 2000* is the development of achievement standards for college students in the United States. This task is not a simple matter of identifying levels of achievement for the nation and tracking them over time to ensure that higher-order skills are acquired by an increasing percentage of students. Standards are not levels of achievement, such as percentile ranks of scaled scores, but

[4]There are two lines of research that seem to have considered the question of higher-order skills is some depth. One has focused its attention on the problem of curricula and teaching strategies to enhance critical thinking and problem solving, in particular. The Delphi Report (cf. Facione, 1990) characterizes this line of work. The other line of research, described in the text, is probably less concerned with specific instructional strategies than it is with the generalizability of any particular strategy. It is important that both lines of research be represented in the evaluation of approaches to assessing higher-order skills. However, because of the comparative lack of clarity regarding definitions of higher-order skills, it seems more important that the psychological research pertaining to them be weighed carefully than would be true if there were a clear curriculum to guide measurement efforts.

rather are value judgments about whether observed levels are acceptable. Needless to say, the success of NACSL as an instrument of federal higher education policy is likely to rest entirely on the degree to which chosen standards are understood and accepted by the public and professional community. The standard setting process is complicated by a number of factors that are somewhat unique to college student assessment. Describing several of these complications is important before outlining any strategy for NACSL that will have standards as a critical component.

Single or Multiple Standards With a system of postsecondary education characterized by pluralism and open access, though with varying degrees of selectivity and admissions standards, a decision must be made regarding the validity of a single achievement standard for all types of institutions. Because colleges and universities in the United States have differences sources of funding and see their particular missions in higher education from sometimes unique perspectives, perspectives known to their applicants, it may be that standards will need to be conditioned on the mission statements of colleges and universities. In other words, NCES may need to develop multiple standards for multiple institutions, particularly if reports of results are disaggregated to the institution, which is the direction NAEP has moved in recent years. On the other hand, if only a single report were issued, for the nation as a whole, then a single standard might make more sense. In either case, the system of achievement standards should be highly sensitive to the nature of the reporting system that NCES develops. This is in the interest of ensuring that the consequences of NACSL do not compromise its validity.

Stability of Standards Regardless of the method that is selected for setting achievement standards, it is critical that they can be defended in terms of their own reliability. Reliability of standards can be defined in a number of ways. For example, if an Angoff or Nedelsky procedure is used that yields specific cut scores for various levels of proficiency, the standard error of measurement at those cut scores can be used to estimate how much chance error is involved in an inference about percentages of a population in the categories on either side of the cut score. This approach directs attention at uncertainty of the inference due specifically to measurement errors in the test materials.

Of equal if not greater importance is gaining an understanding of the standard setting process itself, and quantifying the amount of chance error that exists in the judgments that lead to whatever achievement levels are selected. Replication of a standard setting procedure with an independent sample of judges is an expensive but effective way to evaluate the stability of the process. If an independent sample of experts chooses markedly different achievement levels, then the process of standard setting is untrustworthy with respect to inferences of great social import. A single sample of judges, provided it is large enough, can be used to estimate the chance error of a standard setting process through a jackknife or bootstrap procedure. It will be of utmost importance to establish that the achievement levels selected as NACSL standards are robust to the sampling of judges, and direct estimates of their standard errors should be secured prior to any operational use and reports of results to the public.

III: Programmatic Research and NACSL

The proposed initiative for assessing college student achievement in higher-order cognitive processes involves the breaking of much new ground at the national level in both the content of the assessment and the procedures that might make the entire effort successful. **Because *America 2000* marches in a direction that involves some of the most controversial uses of educational measurement in content domains where much uncertainty exists about construct definitions, NCES should proceed slowly with the development and implementation of an assessment program for NACSL.** This recommendation is not a mere attempt to rain on the *America 2000* parade. The state of knowledge regarding the measurement of higher-order skills at the college level, not to mention the mapping of national achievement scales onto social utility scales, is sufficiently limited that no procedure can be recommended for a national assessment of college students **because of the good data that it has produced** to date. Instead, a program of systematic research directed at the development of procedures and instruments for NACSL should be initiated. A carefully designed research program, predicated on the eventual implementation of a federal system postsecondary measures of achievement, can itself provide preliminary indicators of activities in college classrooms that enhance critical thinking, communication, and problem-solving. The remainder of this section discusses aspects of a federal assessment system for postsecondary education that might be the focus of research activities pursuant to NACSL.

Specification of Content Domains It was argued earlier in this paper that a meaningful assessment, one that truly responds to Goal 5, must gather

information about higher-order skills within the context of specific disciplines. Of interest is not the content of the chemistry or sociology curricula, but rather instances in which evidence of critical thinking, communication, and problem solving is transparent from that content. The development of assessment tasks of the sort envisioned will require many subject matter specialists to think about their expertise in new ways. Several disciplines should be targeted for research projects that contribute to the specification of a content domain for higher-order skills in those disciplines. The collection of work samples on the national level, with sampling stratified by type of institution, could contribute to domain specification. All efforts of this sort would be predicated on the principle that **the definition of content domains for NACSL should be empirically grounded** and not determined on the basis of expert judgments. Work samples are the first step.

Authority for Assessment Initiatives and Results Heated discussions have been in progress around the United States regarding the extent to which assessment activities aimed at creating national standards of achievement should be in the hands of a central authority in Washington. Measurement specialists have long argued that the major value of assessment information is its use at the local level to improve instruction and learning in classrooms. *America 2000* clearly calls for accountability with a national focus. Yet, it would be quite reasonable to suspect that a national report on educational progress, published once a year, would be perceived as remote and even tangential to the important learning experiences of students on American college campuses and have little impact at the local level.

Experience with testing in the public schools indicates that the decentralization of authority has a major impact on the credibility and utility of assessment results. Decentralization in the present context poses major technical problems if the system developed for NACSL must provide aggregate results that can be examined for trends over time. Can some of the models being considered for cluster examinations within the context of the New Standards Project be used at the postsecondary level? This and other questions of policy and technique are ultimately influenced by the kind of reporting needed to satisfy Goal 5. An in-depth policy analysis of the proper federal role in the assessment of college student learning is needed in order to shed light on the degree of centralization that is desirable in NACSL. Research of this sort should be part of the foundation of the federal effort.

Instrument Development and Evaluation Whether the instruments eventually used for NACSL are traditional objective tests or performance-based exercises (or some combination), considerable effort will be needed to ensure their internal validity. Just as the definitions of content domains should be empirically grounded, so too should be any decision regarding the types of tasks that can validly measure the target skills. ED should consider the question of format of the assessment empirically, and initiate research to develop a wide range of potential assessment tasks, from the highly structured to the unstructured and perhaps even ill-structured. Potential assessment tasks should be field tested in a way that provides data to inform decisions related to feasibility of various approaches, reliability, generalizability to the domain or to

other formats, fairness to minorities and other demographic groups of concern, and fidelity to the skills that are targeted by the assessment. Traditional item analyses will be required, as well as judgmental analyses of the content of the tasks, reliability and generalizability analyses of the data from purportedly parallel tasks, and correlational analyses to assess convergent and discriminant validity. These efforts should be coupled with a systematic attempt to develop indicators of college student achievement in higher-order skills that are based on existing sources of information within institutions, and to evaluate their feasibility for NACSL.

Performance Standards Many measurement specialists would question the feasibility of setting standards for college student performance that would be acceptable to all of the audiences of the assessment. There is a pressing need for field studies of the standard setting process and objective evaluation of their results for higher education. As stated previously, such studies should be aimed at understanding the kinds of errors that can obtain from the standard setting process.

Chance errors are important. However, it is equally important to evaluate the representativeness of the judges who can be assembled for standard setting sessions and to assess their ability to gauge achievement for groups they may not have first-hand experience with. Are all of the groups who policymakers determine should play a role in setting achievement standards capable of rendering consistent judgments? Business leaders, for example, are recognized as an important audience for assessment activities at the national

level; however, it is an empirical question as to whether or not they can examine measurement instruments and make consistent decisions about how well college graduates ought to perform on them. The same question could be asked of college professors, subject matter specialists, or political leaders. The appropriate population for the group who will set the standards needs to be carefully considered, and their skill in rendering consistent judgments should be examined empirically.

An additional area of uncertainty in the application of standard setting to college student populations concerns the methodology used. Standard setting methods are known to give widely discrepant results even when applied under the same conditions to the same content. Methods developed by Angoff (1971), Ebel (1956), Nedelsky (1954), and Jaeger (1978) have been used across the United States in certification and competency testing programs. Jaeger (1989) summarizes the substantial variation within testing programs due to the method used to set standards. Of great value to NACSL would be research that examined standard setting procedures with high-level, domain-specific examination materials. Most standard setting activity in testing programs has been concerned with establishing minimum standards in basic and applied skills, or subject matter derived directly from professional training programs, for example, in nursing, teacher certification, or other licensure situations. Experience has shown that standard setting can be a much more difficult task when the focus is placed on higher-order skills, such as problem-solving and critical thinking. Because NACSL is primarily concerned with such skills,

research on the effectiveness of various judgmental methods for establishing levels of proficiency is needed.

Student Motivation Very little is known about the amount of effort a student will expend on a task that is used for external assessment. Results from NAEP do indicate that the problem of student motivation gets worse as students get older. No-shows are more frequent among 17-year olds than they are among 9-year olds. Unscorable essays (those that are blank or that don't respond to the assigned topic) are more frequent among older examinees. To say that little is known is not to ignore a certain amount of common sense, that we can expect motivation to be a serious problem among college students and likely to be more serious when the assessment tasks require that students engage in higher thinking processes. What we don't have a clear understanding of is whether lack of motivation to perform well compromises the results of a national assessment to such an extent that it severely underestimates the quality of schooling. If performance is to be judged "good enough" according to some standard, it is especially important the the effects of motivational influences on performance be measured. There are likely to be data on this issue at the institution level.[5] Research projects that investigate ways of increasing student motivation are an important component of the development effort for NACSL.

[5]Graham (1988) discusses motivation among college students as it pertains to assessment activities, and reviews methods to measure various aspect of student motivation. It would be well to consider attempts at assessing students' motivation to perform on examinations that have no direct bearing on their success in college.

Outlook on Validation The most critical aspect of the development of a national assessment system for higher education is the provision made for ongoing validation studies. The APA/AERA/NCME *Standards for Educational and Psychological Testing* require that both developers and users of instruments establish that assessment information is used in ways that are appropriate given the psychometric characteristics of the measures. In an open market for test materials, developers who can produce the strongest evidence in favor of reliability and validity are in the best position to convince potential users of the value of their instruments. Particularly in an area where content arguments for validity are on soft ground, ED should consider the process of validation as an integral part of the evolution of a national assessment system for higher education. There is much research in outcomes assessment for higher education that demonstrates the extreme difficulty of ensuring validity in the large-scale, high-stakes use of tests to evaluate programs and institutions. Invalidity stems from the limitations of instruments, misconceptions about test use, and general lack of experience and knowledge about measurement. Part of the preliminary effort in designing NACSL should focus on a plan for ongoing validation activities, so that the range of appropriate uses of results from NACSL can be understood when findings are reported to the public.

IV: Implementation of NACSL

Recommendations for further research aren't likely to satisfy those charged with carrying out a policy initiative at the federal level. What is being recommended here is not basic research of the sort that might lead to

implications for a national assessment of college students and might not. On the contrary, the research program being advocated is exactly the sort of thing one does prior to the development of educational measures and all of the efforts are pursued precisely because they contribute knowledge and materials directly to the assessment effort. Thus, ED may want to consider the proposed research agenda itself as the first phase of a larger effort at national assessment. The main purpose behind emphasis on this first phase is simply that, from the standpoint of measurement practice, there are too many uncertainties in the proposed activities to recommend that a particular approach will lead to a valuable assessment, worthy of the public funds that will be needed to support the system in the future.

In 1966, Ralph Tyler wrote of the need for a base of information that could help policy specialists make decisions about public education. Few would say that his words some 25 years ago have become dated:

> The great educational tasks we now face require many more resources than have thus far been available, resources which should be wisely used to produce the maximum effect in extending educational opportunity and raising the level of education. To make these decisions, dependable information about the progress of education is essential. . . . Yet we do not have the necessary comprehensive and dependable data; instead, personal views, distorted reports, and journalistic impressions are the sources of public opinion. This situation will be corrected only by a careful, consistent effort to obtain data to provide sound evidence about the progress of American Education (Tyler, 1966; p. 95).

The national assessment that grew out of Tyler's vision had an open framework, one in which experimentation with instrument design was

encouraged at the expense of building measures that could maintain comparable scales over time. This was because the assessment tasks and exercises themselves were viewed as the primary units of analysis in NAEP's early years, and one of the few aspects of the assessment system that policymakers and the public would be able to understand without special training. In order to be convincing, it was thought that the exercises used in the assessment should have intrinsic value in a way that was apparent to anyone looking at them. Reports published exercises and simple summaries of the percentages of students that could perform them successfully. And the sampling of content domains was broad, with a total of ten hours of exercises used in a matrix sampling design to yield aggregate measures of educational progress in a given content domain.

As noted by Linn (1990), NAEP has changed considerably since Tyler's early recommendations, with the majority of the changes motivated by a perception that NAEP itself should become more of a policy arm for federal evaluation of the state of American education. As an indication of one effect of this change, Linn notes that the 1990 assessment in mathematics for grade 8 had provisions for a total of only one hour and 45 minutes of assessment exercises compared to the ten that Tyler's model allowed. With the need to track achievement over time, and to provide data that can be directly interpretable in terms of educational policy, NAEP has been compelled to narrow its domain definitions to only those about which consensus can be reached. Linn quotes Stake (1971) as an early critic of the consensus process. At a time early in its history, when some were pushing NAEP in the direction it ultimately went, Stake

contended that "the decision to filter all objectives through a committee of subject-matter experts, a committee of educators, and a committee of citizens yields a product that even an ulcer-ridden public can find inoffensive" (1971, p. 58). This perspective gives one good reason to pause and consider the wisdom of moving quickly to establish guidelines for the development of a census assessment for higher education in the United States and whether such an approach would be truly responsive to the public that seeks information or truly consistent with the pluralistic system of higher education that exists in American society.

The research program discussed in this paper can be viewed as a national assessment of sorts, one patterned more after the original NAEP than its descendants, if ED decides to pursue a developmental strategy in which NACSL assumes the role of information provider, separating measurement and public policy. Such a separation would disambiguate findings that are drawn from the actual measurements undertaken in NACSL from the policy statements that ED chooses to make on the basis of NACSL findings. The need for a stronger research foundation for educational reform efforts for the next decade and beyond is well established (cf. Kirst, James & Shulman, 1991).

As implied by the first section of this paper, such an approach quite probably fails to respond in any direct way to Goal 5 of *America 2000*. However, many of the controversies that have surrounded recent attempts to make the results of national assessment relevant in the eyes of politicians and the public stem from attempts to make assessment serve conflicting masters. Efforts

directed at national assessment of college student learning should, from their very inception, create the richest possible source of dependable data about activities in college classrooms and the diverse outcomes of higher education in the United States. Re-creating a national assessment that has innovation through research and experimentation as a central theme would place ED in a much better position to effect the improvements in higher education in the United States that are called for by *America 2000*.

References

America 2000: An education strategy (1991). Washington, D.C.: U.S. Department of Education.

AERA/APA/NCME (1985). *Standards for Educational and Psychological Testing*. Washington, D.C.: American Psychological Association.

Angoff, W. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.) *Educational Measurement*. (2nd. ed., pp. 508-600). Washington, D.C.: American Council on Education.

Baird, L. L. (1988). Diverse and subtle arts: Assessing the generic outcomes of higher education. In C. Adelman (Ed.), *Performance and judgment: Essays on principles and practice in the assessment of college student learning*. Washington, D.C.: U.S. Department of Education.

Cannell, J. J. (1988). Nationally normed elementary achievement testing in America's public schools: How all 50 states are above the national average. *Educational Measurement: Issues and Practices, 7*, 5-9.

Dunbar, S. B., Koretz, D. & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4*, in press.

Ebel, R. L. (1956). Obtaining and reporting evidence on content validity. *Educational and Psychological Measurement, 16*, 294-304.

Facione, P. A. (1990). *Critical thinking,: A statement of expert consensus for purposes of educational assessment and instruction*. Millbrae, CA: California Academic Press.

Gallup, A. M. and Elam, S. M. (1988). The 20th annual Gallup Poll of the public's attitudes toward the public schools. *Phi Delta Kappan, 69*, 33-46.

Glass, G. V. (1978). Standards and criteria. *Journal of Educational Measurement, 15*, 237-261.

Graham, S. (1988). Indicators of motivation in college students. In C. Adelman (Ed.), *Performance and judgment: Essays on principles and practice in the assessment of college student learning*. Washington, D.C.: U.S. Department of Education.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 485-514). New York: Macmillan.

Kirst, M., James, T. & Shulman, L. (1991). Forging a national agenda in educational research. *Education Week, 11* (1), 44.

Koretz, D. M., Linn, R. L., Dunbar, S. B. & Shepard, L. A. (1991, April). *The effects of high-stakes testing on achievement: Preliminary findings about generalization across tests.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Larkin, J. H. (1989). What kind of knowledge transfers? In L. B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser.* Hillsdale, NJ: Erlbaum.

Linn, R. L. (1990, October). *Historical origins and issues in the National Assessment of Educational Progress.* Paper presented at the Institute for Practice and Research in Education forum on Assessment at the National Level, Pittsburgh, PA.

Linn, R. L., Graue, M. E. and Sanders, N. M. (1990). Comparing state and district test results to national norms: The validity of the claims that 'Everyone is above average'. *Educational Measurement: Issues and Practices, 9,* 5-14.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Nedelsky, I. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 14,* 3-19.

Resnick, L. B. (1989). *Knowing, learning, and instruction: Essays in honor of Robert Glaser.* Hillsdale, NJ: Erlbaum.

Rothman, R. (1991). NAEP board fires researchers critical of standards process. *Education Week, 11* (1), 36.

Snow, R. E. & Lohman, D. L. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 263-331). New York: Macmillan.

Stake, R. E. (1971). National assessment. *Proceedings of the 1970 Invitational Conference on Testing Problems: The Promise and Perils of Educational Information Systems* (pp. 53-66). Princeton, NJ: Educational Testing Service.

Thurstone, L.L. (1959). *The measurement of values.* Chicago: University of Chicago Press.

Tyler, R. W. (1966). The development of instruments for assessing educational progress. *Proceedings of the 1965 Invitational Conference on Testing Problems* (pp. 95-105). Princeton, NJ: Educational Testing Service.

Shepard, L. A. (1990). Inflated test score gains: Is the problem old norms or teaching to the test? *Educational Measurement: Issues and Practices, 9*, 15-22.

Yarbrough, D. B. (1991). *Assessing cognitive general education outcomes: Conclusions from a decade of research on the ACT COMP measures* (Appendix E: Supplement to the COMP Technical Report). Iowa City, IA: American College Testing Program.

Fiorello H. LaGuardia Community College THE CITY UNIVERSITY OF NEW YORK
31-10 THOMSON AVENUE, LONG ISLAND CITY, N.Y. 11101 • Telephone (718) 482-7200

John Chaffee, Ph.D.
Director, Creative and Critical Thinking Studies
LaGuardia Community College, The City University of New York

Review of the position paper: *On the Development of a National Assessment of College Student Learning: Measurement Policy and Practice in Perspective*
Stephen B. Dunbar, The University of Iowa

1. What abilities (critical thinking, communication, problem-solving) have been identified and why were they selected?

The author has not identified any specific abilities to be included in a national assessment initiative. Rather, he explains in some detail the obstacles in such a project. "The fact that the goal of interest concerns skills that are not directly taught, coupled with the fact that the ultimate scale of interest is based on social utility and not on achievement *per se*, presents some measurement problems that do not exist in other assessment of educational progress that have been used at local, state, and national levels. Achievement that are, by their very nature, meaningful to the public in the context of a real-life setting do not lend themselves well to observation and objective measurement. Moreover, surrogates for them are prone to charges of contrivance, and the data from them don't generalize well to the criterion situations of real interest."

2. Are the selected abilities appropriate in meeting the goals of this project?

Since the author did not specify abilities to be assessed, this question cannot be addressed.

3. Are the abilities defined in a way that would make possible assessing their development in college students?

Since the author did not define abilities to be assessed, this question cannot be addressed.

4. Do the proposed assessment methods allow for: accurately measuring the abilities; determining the acquisition barriers; identifying effective learning environments?

The author believes that any attempt to construct a national assessment of higher-order thinking and communication abilities is premature. Instead, a preliminary research project should be initiated in order to develop the procedures and instruments required by a national assessment. "The state of knowledge regarding the measurement of higher-order skills at the college level, not to mention the mapping of national achievement scales, is sufficiently limited that no procedure can be recommended for a national assessment of college students *because of the good data that it has produced* to date. Instead, a program of systematic research directed at the development of procedures and instruments of NACSL (national assessment of college student learning) should be initiated. A carefully designed research program, predicated on the eventual implementation of a federal system postsecondary measures of achievement, can itself provide preliminary indicators of activities in college classrooms that enhance critical thinking, communication, and problem-solving."

5. Are the methods or suggestions presented practical, replicable and complete.

The author discusses a variety issues relevant to the research initiative that he is proposing.

6. General Comments:

Conclusion: This paper presents a well-reasoned analysis of the challenges to creating a national assessment of higher-order thinking and communication abilities. To meet these challenges, the author proposes a two-stage process. The first stage would involve research and analysis; the second stage would involve creating assessment instruments and procedures based on the results of the first stage. The author contends that this is the standard approach used for developing educational measures, and that this approach is particularly necessary in the current project because "there are too many uncertainties in the proposed activities to recommend that a particular approach will lead to a valuable assessment, worthy of the public funds that will be needed to support the system in the future."

In my opinion, the author has made a compelling case for this approach, providing that the initial research phase is accomplished in a reasonable, but clearly defined period of time. This will help insure that the assessment stage is initiated in a timely fashion.

Comments on the Position Paper
by Stephen B. Dunbar

Reviewed by: Norman Frederiksen

I should raise one question at the outset. Objective 5 of Goal 5 in the "National Goals for Education" booklet refers to college graduates, while Dunbar refers to assessments at the college student level; I have seen nothing in the paper that refers to graduates. Dunbar has apparently chosen to ignore the graduate term, which in probably a good idea. There are obvious disadvantages to working with college graduates; they will have departed to scattered towns and cities, and few might show up for assessments,

(P.1) The author states that we are "faced vith an initiative that poses some extremely difficult questions...some of which go well beyond any experiences educators in the United States have had in the arena of assessment and public policy." From one point of view this in exaggerated; NAEP and Westat are two organizations that work tosether and could make valuable contributions to the National Goals project (see "The NAEF Guide" by Ina Mullis and her many colleagues).

From another point of view, the assessment task will be very difficult. One reason in that the Goal 5-Objeetive 5 demands are extremely generaILzed as compared with solving a cocrete problem in mathematics or chemistry. Thinking critically, communicating effectively, and solving a problem without a context require quite different knowledge and skills than a problem in mathematics or

chemistry.

(P.3) Dunbar states that Goal 5-Objective 5 (let's call them "5.5" from here on), "are generally understood among educated parsons as generalized outcomes of higher education." However, Baird is cited in a footnote for a different point of view: "There is...a growing literature in cognitive psychology that indicates that we know much more about how such skills operate within specific disciplines than we do about how they transfer across disciplines or how they function... in their own right." He is correct. It seems likely that it will be necessary to work with selected disciples such as mathematics or chemistry. The skill and knowledge required in such a domain are acquired through a great deal of practice, and individuals typically acquire skills in a number of domains, some large and some small. But domains as large and as vague as is implied by 5.5 cannot represent domains in general. The assessments required for the National Goals project should, in my opinion, be accomplished for each of several reelected kinds of performance, not for 5.5 as a whole. Thinking critically about coaching a basketball team in a domain that has very little in common with thinking critically about solving an algebra problem.

(P.4) Dunbar states that "there are actually two scales of interest that need to be defined for NACSL. The first in an achievement scales, and the second ir a utility scales, and the inferences that truly respond to the goal statement of America 2000 require that the achievement scale be mapped onto tho utility scale." I don't understand why the achievement scale should be

mapped on the utility scales; the opposite makes more sense, if there is any need for a utility scale. I don't understand why there should be a utility scale at all, nor do I understand why using 5.5 for assessing gain would work. If any panel is to be involved, I would suggest NAGB.

(P.7) I an interested in the statement in bold type at the bottom of the page: "All of the arguments in this section are based on the premise that the criterion situations in response (to 5.5) involve the target skills in context... " The term in context puzzles me, although it somehow seems to join 5.5 with performance in some domain. Perhaps this could be rewritten so that people like me can understand it. (PP-8-9) It seems to be implied that the only way to demonstrate the validity of a test is "consequential validity": does the use of a test fulfill its desired consequences? If we are talking about NASCL, it is difficult to predict a consequence because we have no notion as to the nature of the test--will it assess "generalized" skill and knowledge, or will it assess skill and knowledge separately in such domains as math, science, literature, or sociology?

One possible consequence in that assessments are made periodically, may once a year, and the desired consequences are changes--did students' performance improve year after year, stay the same, or got worse? Another possibility in that the assessments have instructional value, and the desired consequence in learning by taking tests. Teaching for the test might accomplish this, and was the case when minimum competancy tests were widely used, especially in high-stakes states.

(P.12) Dunbar notes that "the content domains defined [in 5.5] are difficult to operationalize for purposes of measurement," and he is certainly right; on the next page he points out that "there in still a body of research in cognitive psychology that portends difficulty with respect to validity." The problem has to do with 'the degree of transfer that can be expected from the general assessments of higher-order skills to particular applications. The degree of transfer is indeed small, as Larkin makes clear. Tests in such domains as math, science, economics, etc. are <u>not</u> transferable to other such domains; and the skills of 5.5, if any, certainly do not transfer to domain-specific problems.

(Pp.17-18) Dunbar seems to be struggling between two views: that NACSL should be based on (1) highly generalized concepts [5.5] or on (2) the various disciplines that are taught in colleges. He has stated that a meaningful assessment, one that truly responds to 5.5. must obtain infomation about the higher-order skills that are employed in the context of specific disciplines. His interest is not in the content of chemistry or of American history, but on instances in which evidence of critical thinking, communication, and problem solving is "transparent from the content."

Such a piecemeal division of 5.5 might introduce a serious "bias" in learning. One doesn't teach physics by starting with 5.5, and there in much to learn before a physics domain has developed. I doubt that that domain ever becomes a part of 5.5.

My view is that America 2000 should not attempt to develop NACSL tests based on the three skills that make up 5.5.

(P.19) The NACSL should not take the form of traditional objective tests; they should be performance tests that take the form of simulations of real-life situations and that require the student to provide the best answer or answers he can in a free-response format. The situations should of course be aimed at eliciting responses having to do with domain that are typically taught in colleges. Methods for scoring can be developed to suit the responses, in whatever form they may take. Scoring should be done by judges who are experienced in evaluating tent responses, and they should not be based on 5.5. Example: Frederiksen & Ward (1978); Ward, Fredericksen, & Carlson (1980). They should not be aimed specifically at one or more of the 5.5 possibilities.

(P.18) In my opinion, activities related to the National Goals for Educatica should not be in the hands of authorities in Washington, nor should results be reported only at the national level. A relationship similar to that of NAEP and NAGB would be more reasonable, and there in no reason not to report results at the state level.

Some General Comments

Improvement of instruction. in his presidential address George Bush described a set of educational goals that should be accomplished by the year 2000. But no goals were made for improvement in the quality of instruction. The nearest references

to quality of instruction were Goal 4, Objective 2, "The number of teachers with a ... background in mathematics and science vill increase by 50 percent," and Goal 5, Objective 5, "The number of quality programs will increase substantially." The goal-objective 5.5 that was assigned to us does not include anything about improving instruction; Dunbar's paper is concerned almost entirely with the assessment methods defined In 5.5. It would seem that as large a project as this on assessment should be accompanied by efforts to improve instruction. There is an abundance of journal articles and books on improving instruction, and the federal government is supporting much research on the topic. Perhaps the federal government should spend more of its money on improving instruction and less on assessment of the kind described in Dunbar's position paper.

Assessment Methods. If one is to develop methods to assess increases in skill and knowledge, it would be best to work with the disciplines taught in colleges, not 5.5. The test format should not be of the multiple choice variety; instead, it would be best to create tests that are realistic simulations of situations that elicit relevant bahaviors. Problems so posed could increase in difficulty as the college years go by, with increasing need for "higher-order thinking" skills. Suman Chipman has a chapter on what higher-order thinking skills are and how they might be taught (Chapman, S. F. (1990): The higher-order cognitive skills: What they are and how they might be transmitted. In T. C. Sticht, B. A. McDonald, & E. M. J. Beeler (Eds.), The intergenerational transfer of cognitive skills, Norwood, N.J.: Ablex).

An example of a test for assessing higher-order skills in called Formulating Hypotheses, which simulates problem situations like those faced by a behavioral scientint who conceives of and plans research and interprets research data. Each problem requires the subject to read a brief description of an experiment, to study a graph or a table showing the results, and to write hypotheses that might account for the finding. Scoring the test makes use of a classification of the ideas actually written by a group of subjects. The categories in the classification are then evaluated by expert judges: scorers match each subjects' hypothesis to one of the categories, and the score of a subject would be based on the values assigned to the hypotheses (see Frederiksen & Ward (1978): Measures for the study of creativity in scientific problem solving, Applied Psychological Measurement, 2., 1-2; and Ward, Frederikson, & Carlson (1980): Construct validity of free-response and machine-scorable forms of a test, Journal of Educational Measurement, I7, 1129). This article shows that free-response and multiple-choice versions of the test differ greatly with respect to construct validity.

Review of Dunbar's Paper "On the Development of a National Assessment of College Student Learning: Measurement Policy and Practice in Perspective"

by

Ronald K. Hambleton
University of Massachusetts at Amherst

I believe that the National Assessment of Educational Progress (NAEP) provides an excellent framework for addressing national questions about college student proficiency in the areas of higher-order thinking, problem-solving, and communication skills. NAEP, which utilizes (1) national curriculum committees to identify the relevant competencies to assess, (2) measurement committees to concern themselves with questions of valid assessments and scaling, (3) complex sampling designs for selecting participants, and (4) thoughtful and comprehensive reporting methods, has been functioning successfully for over 20 years. In fact, NAEP at the elementary and secondary school levels will provide key information for addressing Goal 3 of America 2000. Clearly, a testing framework is in place that has proven to be successful at providing policy-makers and educators with useful data for decision-making.

Professor Dunbar provides an excellent review of the measurement challenges that are likely to face researchers who attempt to use the NAEP framework (or any framework) in pursuing the assessment of higher-order thinking, problem-solving, and communication skills. Experiences with NAEP will be valuable, but many new measurement challenges will surface when attempts are made to implement NAEP at the college-level.

Professor Dunbar addresses thoughtfully three parts of the current initiative to address Objective 5 of Goal 5 of the educational reform strategy outlined in the America 2000 report

The three parts are:

1. Goals, scales and measurements
2. Measurement issues
3. Programmatic research

With respect to Part 1, I completely agree with Professor Dunbar's concerns about operationalizing what the government means by terms such as "ability to think critically," "communicate effectively," or "solve problems." Often, these "generalized outcomes" are not outcomes that can be pointed to in college curricula, nor do they mean the same thing across disciplines, or even across colleges within the same discipline. Reaching agreement about the meaning will be a major challenge in higher education. Of course, too, the goal itself draws attention to a focus on change. The measurement of change will be hard enough, but what will be the "meaning" of change? Professor Dunbar notes that a type of "social utility" scale will be needed which will be meaningful to the public and not so contrived that generalizations from the assessment tasks to situations of real interest will be limited. Professor Dunbar has put his finger on two very difficult problems to solve.

With respect to Part 2, Professor Dunbar has again provided many important insights. In fact, Professor Dunbar's summary of the consequential validity of many large-scale national testing programs is especially informative, though, surprisingly, he does not include any observations about NAEP, an assessment system most like the one that could be implemented in this project. For the record, I believe NAEP, generally, has been well-received by policy-makers and educators, though, because students, schools, districts, and states are not specifically identified in the NAEP, it is probably viewed by those participating as a low-stakes assessment. But perhaps there is a message here as well. To paraphrase a comment I heard recently, the more high

stakes a testing program is, the less valid the results. If you want valid information about what students have learned, use low-stakes tests, which schools don't view as important! Of course, the problem remains to insure students are motivated to show what they know and can do.

Professor Dunbar goes on (pp. 13-14) to make the very important observation that colleges and universities could agree on a definition of critical thinking, and measurement specialists could produce tasks to measure critical thinking. But, as he notes, it is still possible that performance on these tasks would not generalize (predict) performance on tasks in the workplace. I'm sure, however, that correlations would not be zero, and while the predictions may not be as high as desirable or might be expected, it remains to be seen whether they would be of a size to justify the usefulness of the critical thinking tests.

Professor Dunbar goes on in his paper to address standards. Single or multiple standards, and the stability of standards which are addressed in the paper are certainly important concerns. That is, do you set different standards to accommodate the particular missions of colleges or universities and report the results separately? Would different groups of judges working from the same guidelines and specifications produce similar standards? In any case, the National Assessment Governing Board (NAGB) recently set standards on the 1990 NAEP Mathematics Assessment. Many principles were learned about the selection and training of judges, data analysis, and other aspects of the standard-setting process that could be useful in future national standard-setting initiatives. Suffice to say here, the NAGB effort was expensive, time-consuming, and many measurement problems arose in the process. We note in passing that standard-setting in mathematics seems considerably easier than

in the area of critical thinking with its complex definitions, ambiguities, performance assessments, and arbitrary scoring methods.

Professor Dunbar addresses a research agenda in Part 3 of his paper. I want to support completely Professor Dunbar's recommendations. Yes, we should move slowly because we are without a strong database. We neither know very much about the skills of importance nor how they might be reliably and validly measured within a national assessment.

In summary, I believe that Professor Dunbar has provided a great service to the government and the workshop participants by preparing a clear, insightful presentation of many of the measurement problems and possible solutions associated with a project to assess the development of higher-order thinking skills among college students. My own list of problems includes (1) the development of measurement scales for reporting results (will they need to be multi-dimensional?), (2) standard-setting (a difficult problem that perhaps can be avoided if the focus shifts to measuring growth rather than determining the percent of students who are doing well enough), (3) construction of performance assessments (which seems essential for valid assessment) and all of the problems associated with performance testing (e.g., low task generalizability, inter-rater reliability problems, high costs, scoring problems), and (4) problems of motivating students to take any national assessment of critical thinking skills seriously.